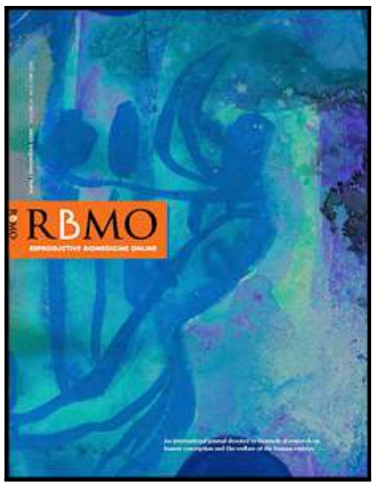
Journal Pre-proof



Assessing the performance of generative AI chatbots in preimplantation genetic testing: a comparative study of expert evaluations

Belén Lledo , Paola Carbone , Jose A. Ortiz , Ruth Morales , Adoración Rodríguez-Arnedo , Leyre Herrero , Elisa Alvarez , Jorge Ten , Lydia Luque , Juan C. Castillo , Jordi Suñol , Annalisa Racca , Andrea Bernabeu

PII: S1472-6483(25)00482-1
DOI: https://doi.org/10.1016/j.rbmo.2025.105275
Reference: RBMO 105275

To appear in: *Reproductive BioMedicine Online*

Received date: 11 May 2025
Revised date: 17 August 2025
Accepted date: 19 September 2025

Please cite this article as: Belén Lledo , Paola Carbone , Jose A. Ortiz , Ruth Morales , Adoración Rodríguez-Arnedo , Leyre Herrero , Elisa Alvarez , Jorge Ten , Lydia Luque , Juan C. Castillo , Jordi Suñol , Annalisa Racca , Andrea Bernabeu , Assessing the performance of generative AI chatbots in preimplantation genetic testing: a comparative study of expert evaluations, *Reproductive BioMedicine Online* (2025), doi: https://doi.org/10.1016/j.rbmo.2025.105275

Assessing the performance of generative AI chatbots in preimplantation genetic testing: a comparative study of expert evaluations

Belén Lledo[a,*], Paola Carbone[a], Jose A. Ortiz[a], Ruth Morales[a], Adoración Rodríguez-Arnedo[b] , Leyre Herrero[b], Elisa Alvarez[b], Jorge Ten[b], Lydia Luque[c], Juan C. Castillo[c], Jordi Suñol[c], Annalisa Racca[c], Andrea Bernabeu[c,d]

a Instituto Bernabeu Biotech, 03016, Alicante, Spain

b Reproductive Biology. Instituto Bernabeu of Fertility and Gynaecology, 03016, Alicante, Spain

c Reproductive Medicine. Instituto Bernabeu of Fertility and Gynaecology, 03016, Alicante, Spain

d Cátedra de Medicina Comunitaria y Salud Reproductiva, Miguel Hernández University, Alicante, Spain

*Corresponding author. E-mail address: blledo@institutobernabeu.com (Belén Lledó)

Funding statement: No funding for the project

Conflict of interest statement for all authors: None

CRediT Authorship Contribution Statement

Belén Lledo: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Writing – original draft, Supervision. Paola Carbone: Investigation, Writing – review & editing. Jose A. Ortiz: Conceptualization, Methodology, Data curation, Formal analysis, Writing – review & editing. Ruth Morales, Adoración Rodríguez-Arnedo, Leyre Herrero, Elisa Alvarez, Jorge Ten, Lydia Luque, Juan C. Castillo, Jordi Suñol and Anna L. Racca : Investigation, Writing – review & editing. Andrea Bernabeu: Funding acquisition, Writing – review & editing

Attestation statements: Data will be made available to the editors of the journal pre and/or post publication for review or query upon request.

Data sharing statement: not applicable

ABSTRACT

Research Question: How reliable are generative artificial intelligence (AI) chatbots in responding to patient-relevant questions about preimplantation genetic testing (PGT), as evaluated by reproductive medicine specialists?

Design: A prospective evaluation was conducted comparing three publicly available generative AI models—ChatGPT-3.5, Gemini-1.5, and Llama-2. Twelve reproductive medicine specialists from different clinics assessed chatbot-generated responses to 13 PGT-related questions, divided into simple and controversial categories. Each response was scored from 0 to 5 using

predefined criteria. Assuming all answers were excellent, the maximum score was 25 points for simple questions and 40 points for controversial ones.

Results: A total of 156 evaluations were completed. Among simple questions, the lowest-rated response was to "What are the types and techniques used for PGT?" (mean score: 2.83±0.94). For controversial questions, "What is the percentage of aneuploidy that allows an embryo to be defined as mosaic?" scored lowest (2.67±1.22). ChatGPT performed best across both categories (simple: 16.83±1.80; controversial: 27.75±4.49), followed by Gemini (14.92±2.02; 26.08±3.99) and Llama (13.58±3.60; 16.92±4.96). Statistically significant differences were observed, particularly between ChatGPT and Llama ($p=0.027$ for simple, $p<0.001$ for controversial), and between Gemini and Llama for controversial questions ($p<0.001$). No significant performance differences were noted across participating specialists.

Conclusions: Generative AI shows moderate reliability in addressing PGT-related inquiries, with ChatGPT and Gemini outperforming Llama. While performance was higher for simple than for controversial questions, the variability underscores the need for clinical oversight. Further refinement and validation are essential before widespread integration of AI tools in reproductive medicine.

Keywords: Generative AI, chatbots, PGT and embryo mosaicism.

INTRODUCTION

Preimplantation genetic testing (PGT) is a diagnostic procedure in ART (Assisted Reproductive Technologies) that involves the genetic evaluation of embryos before implantation. It was initially developed as PGT-M to prevent the transmission of inherited monogenic disorders, and was later extended to PGT-A (Preimplantation genetic testing of aneuploidies) and PGT-SR (Preimplantation genetic testing for structural rearrangements) for the assessment of chromosomal abnormalities in embryos. While the primary objective of PGT-M and PGT-SR is clearly to prevent the transmission of a specific genetic alteration, PGT-A aims to improve overall reproductive outcomes. The goal of PGT-A is to reduce time to pregnancy and lower the risk of miscarriage by optimizing embryo selection (Gudapati et al., 2024). A particular challenge in PGT-A is mosaicism. This phenomenon is common in human embryos and can result mainly from errors during mitosis. Mosaicism complicates the interpretation of PGT-A results, as it may lead to embryos being incorrectly classified. Although mosaic embryos can result in healthy live births, their clinical management remains a topic of debate and requires genetic counseling for patients, along with their understanding of the potential implications (Cheng et al, 2022).

Artificial Intelligence (AI) is a multidisciplinary field of computer science that seeks to develop systems capable of simulating human intelligence through the optimization of mathematical functions. In general terms, AI encompasses systems that can understand natural language,

process visual and auditory inputs, and interact with the environment, often in combination with robotic systems. Its goal is to replicate human intelligence to address complex problems (J. Joosten et al, 2024). A chatbot is a practical application of generative artificial intelligence that serves as a model for human-computer interaction (Bansal Khan, 2018). AI chatbots have shown great potential in fields like healthcare, assisting medical professionals in the diagnostic process (Nassiri and Akhloufi, 2024). Likewise, in the medical sector, chatbots can support diagnostic processes, provide patient education, and facilitate decision-making. AI has also transformed medicine by enabling personalized and precision approaches in healthcare (Kshetri et al., 2023). However, most AI chatbots available for public use are based on a general model and are not trained nor fine-tuned specifically for the medical field.

In the field of reproductive medicine, AI has gained ground, particularly in assisted reproductive technologies. ART generates a large amount of data, making it an ideal candidate for AI algorithm applications (Ortiz et al. 2022). These algorithms can assist in ovarian stimulation personalization, sperm and egg selection, embryo quality evaluation, embryo implantation prediction and DNA screening for fertility issues (Liu et al. 2022). Generative AI chatbots have significant potential in reproductive medicine, particularly in the realm of PGT where the precision and reliability of the information provided are fundamental to ensure that couples receive accurate and scientifically based advice. One of the main challenges in PGT is that patients often face inconsistent or highly technical information that is difficult to interpret, which may hinder informed decision-making. Previous studies have highlighted that patients considering PGT have specific decisional needs and benefit from clear, lay language that is simple to understand yet sufficiently informative to support their choices (Cheng et al., 2022). In this context, generative AI chatbots could offer an innovative solution to improve consistency and accessibility of information. However, their reliability and accuracy must be thoroughly evaluated to ensure they can effectively support clinical decision-making and patient education. Our study aimed to evaluate three AI chatbots, assessed by twelve reproductive medicine specialists, to determine their effectiveness in the field of PGT. This study contributes to understanding the capabilities and limitations of AI in this specialized field, paving the way for its future integration into healthcare practices.

MATERIALS AND METHODS

Study design

A prospective observational study was designed to assess the ability of free, publicly available versions of ChatGPT-3.5 (OpenAI, California, U.S.), Gemini-1.5 (Google DeepMind, London, U.K.), and Llama-2 (Meta AI, Menlo Park, U.S.) chatbots to provide appropriate answers to

patients asking questions about PGT and embryo mosaicism. The design encompasses several fundamental phases, including the selection and categorization of questions, the evaluation of responses, the criteria used, and the scoring scheme employed.

Selection and categorization of questions

Geneticists in reproductive medicine curated 13 patient-relevant PGT questions. The process of curation included referencing from established resources namely, the European Society of Human Reproduction and Embryology (ESHRE) and Preimplantation Genetic Diagnosis International Society (PDGIS). The queries presented to all chatbots were standardized and meticulously categorized to cover a wide range of pertinent situations and information within this field of study. The identified categories included: (1) Simple and Direct Questions (SQ): These questions are formulated to evaluate the reliability of AI chatbots at a superficial level and (2) Controversial Questions (CQ): These questions are derived from the doubts and debates within the scientific community. The full list of the 13 curated questions is shown in supplementary data (Table 1S). The queries were introduced to each publicly available automated intelligence chatbot at the website in their free and open-access subscription between February to April 2024.

Evaluation of responses, criteria used and scoring scheme

To ensure a rigorous and reliable analysis of the responses provided by AI chatbots, it is essential to establish clear and well-defined criteria. The evaluation criteria include: Accuracy (the precision of the information provided in relation to known and verifiable scientific data), completeness (the ability of the response to comprehensively cover all relevant aspects of the posed question), clarity (the transparency and ease of understanding of the response, avoiding ambiguity and unexplained technical terms) and consistency with Scientific Sources (the conformity of the information provided to authoritative and up-to-date scientific sources).

Both the query as well as generated full response from ChatGPT-3.5, Gemini-1.5, and Llama-2 chatbots generated answers were provided to each specialist separately. The answers were critically evaluated and scored by a panel of 12 reproductive medicine specialists who were blinded to each other's evaluations: 4 embryologist, 4 geneticists and 4 gynaecologists. A numerical scale from 1 to 5 were assigned to each answers according to the following:

(1) Inadequate Response. This score indicates that the response contains incorrect or misleading information, lacks coherence with scientific sources, and fails to adequately address the question posed. Such responses are not reliable and do not meet the basic requirements of accuracy and relevance.

(2) Insufficient Response. Responses that receive this score contain partially correct information but are incomplete, lack clarity, and may only partially address the question. These responses show some effort but do not provide a comprehensive or satisfactory answer.

(3) Adequate Response. This score is assigned to responses that are correct but not exhaustive. The information provided is accurate and relevant but may lack depth or detailed coverage of all aspects of the question. Clarity is sufficient, and the response meets the basic expectations but does not excel.

(4) Good Response. Responses that receive this score are generally complete and accurate. They cover most of the relevant aspects of the question with good clarity and coherence. The information is well-organized, and the response demonstrates a solid understanding of the topic, aligning well with scientific sources.

(5) Excellent Response. This highest score indicates a response that is fully accurate, comprehensive, and exceptionally clear. The response addresses all relevant aspects of the question in detail, is well-articulated, and demonstrates a deep understanding of the topic. It is perfectly coherent with scientific sources and sets a high standard for reliability and quality.

Statistical analysis

Data analysis was performed using R software version 4.3.1. A descriptive analysis of the scores assigned to chatbot responses was conducted, presenting continuous variables as means and standard deviations (SD). Comparative analysis was performed analyzing each chatbot's mean score and identifying the number of highest-rated replies. To assess differences in performance among the generative AI models (ChatGPT-3.5, Gemini-1.5, and Llama-2), a one-way analysis of variance (ANOVA or Kruskal Wallis according to the distribution of the variables) was performed to compare scores for simple and controversial questions. The Bonferroni multiple comparisons test was used to adjust p-values in post hoc analyses. Differences between specialty groups were analyzed using ANOVA tests and Fisher's exact test for categorical data. Normality tests were conducted to determine the suitability of the applied statistical methods. A p-value < 0.05 was considered statistically significant.

RESULTS

Comparative analysis

There was a total of thirteen questions, twelve expert graders and three chatbots to evaluate given a total number of answers of n=468. Overall, the mean score of the expert panel for simple questions given by the AI chatbots was 3.02 ± 0.88 and for controversial questions 2.95 ± 1.12. The simple and controversial questions with the lowest rates were: "What are the types and techniques used for PGT?" (2.83 ± 0.94) and "What is the percentage of aneuploidy

that allows an embryo to be defined as mosaic?" (2.67±1.22), respectively. This suggests that even the most basic inquiries posed challenges for the AI chatbots, highlighting a potential gap in their knowledge or ability to convey complex scientific information accurately, highlighting areas for further improvement.

Table 1 shows the evaluation of three generative AI chatbots and revealed a significant variance in the performance of the chatbots when addressing both simple and controversial questions and that remained far from the maximum score of 25 points for simple questions and 40 points for controversial ones, assuming all answers were excellent (Table 1). Different performance was found across question types. For SQ, the total score for ChatGPT (16.83±1.80) was significantly higher than those for Gemini (14.92±2.02) and Llama (13.58±3.60) ($p=0.0172$). Notably, ChatGPT outperformed the other chatbots in several individual simple questions, including SQ_1 ($p=0.0222$), SQ_2 ($p<0.001$) and SQ_3 ($p=0.010$). However, some questions (e.g., SQ_4 and SQ_5) showed no significant differences between the chatbots. Regarding the controversial questions (CQ), ChatGPT also scored the highest overall (27.75±4.49), followed by Gemini (26.08±3.99) and Llama (16.92±4.96) ($p<0.001$). Significant differences were observed in several controversial questions, particularly CQ_2 ($p=0.004$), CQ_3 ($p<0.001$), CQ_4 ($p=0.001$), CQ_6 ($p < 0.0012$), and CQ_7 ($p < 0.0012$), where ChatGPT performed notably better than the others. On the other hand, for CQ_8, although was not significant difference a trend had found ($p=0.050$). Pairwise comparisons highlighted that ChatGPT outperformed Llama in both question types ($p=0.027$ SQ and $p<0.001$ for CQ). Gemini also showed superiority over Llama in controversial questions ($p<0.001$), but no difference was observed between ChatGPT and Gemini for either question type. These results showed ChatGPT generally performed better than both Gemini and Llama across both question types, with particularly strong results in controversial questions.

The data presented in the Figure 1 shows the distribution of responses for simple and controversial questions. For simple questions, the results indicate that the highest proportion of responses graded as Adequate response (43.89%) and Good response (29.44%) categories, suggesting an overall acceptable performance. However, a considerable number of responses were classified as Insufficient (19.44%) and Inadequate (5.56%), highlighting areas that may require improvement. Also 1.67% falls within Excellent response. Interestingly, Adequate response exhibited a downward trend, from 50.0% in SQ_1 to 38.9% in SQ_5, which may indicate a decline in response quality as the questionnaire progressed into embryo mosaicism topic, suggesting that these questions may have been more challenging for chatbots.

As for simple questions, for controversial questions the majority of responses fall into the Adequate response (31.25%) and Good response (26.51%) categories and only a 5.9% falls

within Excellent response. Insufficient and Inadequate responses were more prevalent in the controversial than simple questionsP (20.14% and 13.19%, respectively for CQ and 19.44% and 5.56% for SQ), indicating a potential increase in difficulty or lower chatbot performance. Inadequate response was highest in CQ_1 (22.2%), suggesting potential difficulties with that question.

Expert Comparative analysis

Table 2 shows the comparative evaluation of the three chatbots by three reproductive medicine specialties: Geneticists, Embryologists, and Gynecologists. The results for simple questions (SQ) indicate that ChatGPT consistently achieved higher scores compared to Llama, while Gemini performed at an intermediate level in most cases. However, only among Embryologist this difference reaches statistical significant (p=0.024). ChatGPT achieved the highest score 17.75±0.96, followed by Gemini 16.00±1.63 and Llama 13.25±2.06, suggesting superior performance by ChatGPT in this context. For geneticists and gynecologists, ChatGPT obtained

higher score, followed by Gemini and Llama, however, the differences were not statistically significant. For controversial questions different performance was found among the three chatbots by the specialist. ChatGPT obtained the highest scores in almost all categories. For this type of questions among geneticists the difference among the three chatbots was significant different, the total score for ChatGPT was 28.25±2.87, identical to Gemini 28.25 ±4.65, but higher than Llama 16.25±5.12 (p=0.005). For embryologists, a similar behaviour was observed, with ChatGPT scoring 29.25±4.50, Gemini 27.25±2.06, and Llama significantly lower at 14.75±3.77 (p<0.001). On the other hand, for gynecologists, the differences in the performance between the three chatbots were not statistically significant (p=0.302).

Figure 2 shows a comparison of the evaluation of each chatbot across different expert panel. The evaluation of SQ and CQ responses across the three chatbots—ChatGPT, Gemini, and Llama—revealed subtle differences in performance, with no statistically significant variations across the specialist groups. For the SQ, ChatGPT rated similarly across all specialist groups, with the total score being 17.75±0.96 for embryologists, 16.00±1.63 for geneticists, and 16.75±2.50 for gynecologists. Embryologists generally rated ChatGPT the highest, particularly in SQ_1, SQ_2 and SQ_3 where the chatbot scored 3.75. However, these differences did not reach statistical significance. In the CQ, ChatGPT's performance was especially strong among embryologists and geneticists, with a total score of 29.25±4.50 and 28.25±2.87, respectively versus 25.75±6.08 for gynecologist, again without statistical significance. However, notably

embryologists rated ChatGPT significantly higher in CQ_1, where it scored 3.50±1.29, compared to Gemini 2.75±0.5 and Llama 1.50±0.58 (p=0.039).

The performance of Gemini, evaluated by embryologists, geneticists, and gynecologists, was relatively consistent across the groups, especially for the simple questions. For SQ, Gemini's rated higher by the embryologists, who gave it a total score of 16.00±1.63, while geneticists rated it slightly lower at 14.75±1.71, and gynecologists rated it the lowest at 14.00±2.58. However, the differences in total scores were not statistically significant (p=0.412), indicating that all three groups found Gemini to perform similarly. In individual simple questions despite some variability in questions scores, the differences were not statistically significant (p> 0.05). When evaluating the controversial questions, Gemini continued to show comparable performance across the groups of specialists in the total and individual questions score.

Finally, the performance of Llama across the three specialist groups revealed distinct patterns in both simple and controversial question evaluations. For the SQ, Llama received a total score of 13.25±2.06 from embryologists, 11.25±4.99 from geneticists, and 16.25±1.26 from gynaecologists. Notably, gynaecologists gave the highest score overall, while geneticists rated it the lowest. However, the differences in total scores were not statistically significant (p=0.132), suggesting that, despite the variation, there was a strong consensus regarding Llama's performance. In the individual SQ evaluations, no significant differences were found across the groups. For the CQ, Llama's total scores again were highest among gynaecologists 19.75±5.68, with embryologists scored the lowest 14.75±3.77, and geneticists falling in between 16.25± 5.12 but these differences were not statistically significant. This suggests that while there were some variabilities in the individual evaluations, they did not constitute a significant overall trend. In individual CQ evaluations, there was a notable variation in the responses from each group in CQ_5, gynaecologists rated Llama significantly higher 3.50±0.58 than embryologists 2.25±1.26 and geneticists 1.50±0.58 (p=0.0482). This result may suggest that gynaecologists found Llama's responses to this question more satisfactory compared to the other specialists. In contrast, for the rest of CQ questions the variations did not reach statistical significance.

DISCUSSION

This study aimed to evaluate the performance of three commonly free available generative AI chatbots—ChatGPT-3.5, Gemini-1.5, and Llama-2—specifically in the context of answering questions related to Preimplantation Genetic Testing (PGT) and embryonic mosaicism, a key issue in reproductive medicine. The results presented significant variability in the performance of these chatbots. Our results suggest that ChatGPT consistently outperforms the other

models across most question types, although all chatbots demonstrate limitations in providing highly accurate and reliable responses, particularly for complex or controversial topics. The interaction of two key technologies, AI and reproductive medicine, raises fears, needs respect of ethical and legal principles as a means of regulating the industry, public health, and patient rights. From our knowledge this is the first study to evaluate this topic.

The choice of topic has fallen on PGT due to its nature as a subject rife with numerous uncertainties and debates. These debates pertain not only to the reliability of the process or its actual ability to provide benefits but also to the complex decision-making process involved mainly when patients faced with embryo mosaicism (Cheng et al, 2022). In fact, since the first publication describing the birth of healthy children after the transfer of embryos classified as mosaic after PGT-A (Greco et al., 2015), different scientific societies have published guidelines or position statements regarding their recommendations on management and prioritization criteria for couples considering the transfer of a mosaic embryo (Muñoz et al, 2024).

The selection different type of questions had different goals consistent with methodologies used in previous research evaluating AI performance. The aim of simple and direct questions was to assess the chatbots' ability to provide correct and consistent responses to basic and well-defined queries that do not require high processing complexity. These AI chatbots have shown strong performance in answering such queries due to their ability to access vast datasets and provide relatively accurate answers in general domains (Reda, 2024). However, the aim of controversial questions was designed to test the extent to which AI tools can be considered reliable. Controversial questions require chatbots to address complex or debated topics, challenging their ability to handle nuanced information and provide responses that reflect a deep understanding of the scientific and controversies surrounding scientific and ethical issues (Chakraborty et al 2023).

In this study we used a rigorous and reliable score system. A detailed description of the scoring system used is fundamental to ensure consistency and objectivity in the evaluation of responses by different experts. A well-defined scoring system helps to standardize the assessment process, making it easier to compare and interpret results, facilitating a clear and uniform understanding of what each score represents. Using this structured and objective scoring system, we could effectively evaluate the performance of AI tools. This approach allows for the quantification of response quality, making it easier to identify areas of strength and those in need of improvement. Additionally, the adoption of a standardized scale enhances the robustness of the study by enabling the replication of results across different trials and contexts, thereby contributing to the overall reliability and validity of the findings.

Regarding the selection of the three chatbots, each offer distinct capabilities that could influence the generation of responses in specialized scientific contexts such as preimplantation genetic testing (PGT) and embryonic mosaicism. ChatGPT, developed by OpenAI, is an advanced language model based on GPT technology (Roumeliotis and Tselikas., 2023). It is capable of generating human-like text responses and has applications in customer service, education, and healthcare. However, its lack of transparency in decision-making and the absence of empathy

might pose challenges when addressing sensitive genetic topics as PGT and embryo mosaicism which requires precision and emotional sensitivity (Dwivedi et al., 2023). On the other hand, Gemini, another prominent model developed by Google Research and DeepMind, is multimodal, allowing it to process text, images, audio, and video (Imran and Almusharraf, 2024). This makes it particularly valuable for analyzing genetic test results, such as those involving visual data in embryo mosaicism diagnosis, and its accuracy enhances its suitability for evidence-based decision-making in PGT (Wang et al, 2025). However, the complexity of integrating multiple data types might make it less efficient in scenarios where a quick, text-based response is needed. Lastly, Llama, an open-source language model by Meta AI, is valued for its adaptability, enabling customization for specific needs, such as analyzing genetic patterns related to embryonic mosaicism (Touvron et al., 2023). This flexibility allows it to be fine-tuned for highly specialized contexts, but its lack of multimodal capabilities and potential concerns regarding data privacy due to its open-source nature could limit its application in sensitive genetic data. Overall, each model brings unique strengths and potential drawbacks, which should be carefully considered when selecting the most appropriate tool for generating responses in the specialized fields of PGT and embryonic mosaicism.

The three examined chatbots, provided moderately satisfactory results, achieving an acceptable average score in both simple and controversial questions, demonstrating a modest response capability. However, the questions with the lowest performance, highlight the difficulty that AI systems face when addressing questions that require both a precise understanding of biological processes and the ability to present information in a context-sensitive manner (Gignac and Szodorai, 2024). The reduction in response quality, particularly for the topic of mosaicism, supports this notion. As the questions moved towards more controversial topics, chatbot performance decreased, potentially reflecting both the complexity of the subject matter and the controversy that exists within the field of embryonic mosaicism, such as the threshold for defining mosaicism, as evidenced by the larger proportion of inadequate responses. This agrees with previous studies from other medical AI assessments, where models have shown a reduction in accuracy as questions move from straightforward

tasks to those that require reasoning about probabilities or deal with areas of medical controversy (Giannakopoulos et al, 2023; Schmidgall et al, 2024). This indicates that AI, while promising, still requires considerable refinement in addressing controversial topics where scientific consensus is lacking. As for comparative assessment, the overall variability in chatbot performance, ranging from ChatGPT's highest scores to Llama's lowest, reflects the different capabilities of generative AI when applied to specific fields such as genetics and reproductive medicine. In this study, ChatGPT outperformed both Gemini and Llama, in both type of questions, which may point to its more sophisticated underlying model or training on a broader dataset related to science and medicine (Wang et al., 2025). This finding aligns with previous studies. In the field of autoimmune liver diseases ten liver specialists systematically evaluated four chatbots to determine their performance (Daza et al, 2024). Although the authors evaluated different chatbots ChatGPT outperformed Gemini.

About expert evaluation, the professional background of the evaluators, influenced the chatbot scores. The superiority of ChatGPT over Gemini and Llama, was most clearly recognized by embryologists and geneticists, who assigned significantly higher scores to ChatGPT's responses, especially for CQ, compared to those of Llama. Interestingly, the differences among the chatbots were not statistically significant within gynecologists. Notably, these results suggests that the difference in chatbot evaluation across specialties highlight the subjective influence of professional context. Embryologists and geneticist appeared more sensitive to the technical accuracy and relevance of chatbot responses—perhaps reflecting their routine clinical practice with highly specific biological detail. This contrasts with gynecologists, whose evaluations were more uniform across chatbots. Conversely, the comparison of the evaluation of each chatbot across different expert panel revealed subtle differences in performance, with no statistically significant variations across the specialist groups. Although these results may appear to contrast with previous findings, our study showed that gynecologists rated Llama higher than embryologists and geneticists did. This variation aligns with findings from earlier studies that highlight how user expertise and daily clinical focus affected the perceived utility and credibility of AI-generated content (Goodman et al, 2023).

While this study offers valuable insights into the performance of generative AI in the context of preimplantation genetic testing and embryonic mosaicism, several limitations must be acknowledged. One major limitation is the restricted number of chatbots evaluated. Although the selection of ChatGPT, Gemini, and Llama was intentional, representing some of the most widely accessible and technologically distinct free model, the exclusion of other commercially or academically relevant models may limit the generalizability of our findings. This choice was

primarily guided by practical considerations, including the widespread use and availability of these three chatbots at the time of the study, as well as their representativeness in terms of different AI architectures. The independence and standardized phrasing of questions, each asked in a separate session, allowed for unbiased assessment of each chatbot's performance, preventing prior responses from influencing subsequent answers. However, question structure may still impact response accuracy, suggesting a potential area for further investigation. Furthermore, although we implemented a rigorous scoring system and engaged specialists from three relevant fields, the subjective nature of expert evaluation may still introduce bias. Future work incorporating a broader range of models and larger panels of evaluators would enhance the robustness of comparative assessments.

In conclusion, this study demonstrates the different capacities of generative AI chatbots to address both simple and controversial questions in the field of reproductive genetics. ChatGPT-3.5 consistently showed superior performance, particularly when evaluated by embryologists and geneticists, emphasizing its potential utility in medical education or as a supportive tool for patient communication. Gemini-1.5 shows promise as a viable alternative, particularly for moderate-difficulty queries, while Llama-2 may require further optimization for specific applications. However, the overall modest performance across all models for controversial questions highlights a clear limitation in current AI tools when addressing complex biomedical content that lacks consensus or requires ethical discernment. These findings emphasise the need for continued refinement of Large Language Models, especially in tailoring responses to highly specialized domains. Future studies should expand the scope by including more models, incorporating multilingual or multimodal assessments, and testing performance in real-world clinical scenarios. Additionally, efforts to develop specific AI systems trained on curated reproductive medicine literature could help bridge current gaps and improve the reliability of AI in supporting reproductive health professionals.

BIBLIOGRAPHY

Bansal, H and Khan, R. A Review Paper on Human Computer Interaction. International Journal of Advanced Research in Computer Science and Software Engineering. 2018. 8:53. 10.23956/ijarcsse.v8i4.630.

Chakraborty C, Pal S, Bhattacharya M, Dash S, Lee SS. Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science. Front Artif Intell. 2023 Oct 31;6:1237704. doi: 10.3389/frai.2023.1237704. PMID: 38028668; PMCID: PMC10644239.

Cheng L, Meiser B, Kirk E, Kennedy D, Barlow-Stewart K, Kaur R. Decisional needs of patients considering preimplantation genetic testing: a systematic review. Reprod Biomed Online. 2022 May;44(5):839-852. doi: 10.1016/j.rbmo.2021.12.011. Epub 2021 Dec 21. PMID: 35183447.

Cheng L, Meiser B, Kennedy D, Kirk E, Barlow-Stewart K, Kaur R. Exploration of decision-making regarding the transfer of mosaic embryos following preimplantation genetic testing: a qualitative study. Hum Reprod Open. 2022 Aug 22;2022(4):hoac035. doi: 10.1093/hropen/hoac035. PMID: 36157005; PMCID: PMC9492260.

Daza J, Bezerra LS, Santamaría L, Rueda-Esteban R, Bantel H, Girala M, Ebert M, Van Bömmel F, Geier A, Aldana AG, Yau K, Alvares-da-Silva M, Peck-Radosavljevic M, Ridruejo E, Weinmann A, Teufel A. Evaluation of four chatbots in autoimmune liver disease: A comparative analysis. Ann Hepatol. 2024 Aug 13;30(1):101537. doi: 10.1016/j.aohep.2024.101537.

Dwivedi YK, Kshetri N,Hughes L, Slade EL, Jeyaraj A, Kar AK, Baabdullah AM., Koohang A., Raghavan V., Ahuja M., Albanna H., Albashrawi MA., Al-Busaidi AS., Balakrishnan J., Barlette Y, Basu S., Bose I., Brooks L., Buhalis D., Carter L., Wright R. Opinion Paper: "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. 2023. International Journal of Information Management (71):102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.

Gignac GE,and Szodorai ET. Defining intelligence: Bridging the gap between human and artificial perspectives. 2024. Intelligence (104):101832. https://doi.org/10.1016/j.intell.2024.101832.

Giannakopoulos, K., Kavadella, A., Aaqel Salim, A., Stamatopoulos, V. & Kaklamanos, E. G. Evaluation of the performance of generative AI large Language models chatGPT, Google bard, and Microsoft Bing chat in supporting evidence-based dentistry: comparative mixed methods study. J. Med. Internet Res. 25, e51580. https://doi.org/10.2196/51580 (2023).

Goodman RS, Patrinely JR, Stone CA, Zimmerman E., Donald RR., Chang SS., Berkowitz ST., Finn AP., Jahangir E., Scoville EA., Reese TS., Friedman DL., Bastarache JA, van der Heijden YF., Wright JJ., Ye F., Carter N., Alexa MR., Choe JH, Chastain CA., Zic JA., Horst SN., Turker I., Agarwal R., Osmundson E., Idrees K., Kiernan CM., Padmanabhan C., Bailey CE., Schlegel CE., Chambless LB., Gibson MK., Osterman TJ., Wheless LE., Johnson DB., 2023. Accuracy and Reliability of Chatbot Responses to Physician Questions. JAMA Netw Open. 6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483

Greco, E, Minasi, MG, Fiorentino, F Healthy Babies after Intrauterine Transfer of Mosaic Aneuploid Blastocysts. N Engl J Med. 2015; 373:2089-2090 PMID: 26581010

Gudapati S, Chaudhari K, Shrivastava D, Yelne S. Advancements and Applications of Preimplantation Genetic Testing in In Vitro Fertilization: A Comprehensive Review. Cureus. 2024 Mar 31;16(3):e57357. doi: 10.7759/cureus.57357. PMID: 38694414; PMCID: PMC11061269.

Imran M, Almusharraf N. Google Gemini as a next generation AI educational tool: a review of emerging educational technology. 2024 Smart Learn. Environ. 11, 22. https://doi.org/10.1186/s40561-024-00310-z

Joosten J, Bilgram V, Hahn A, Totzek, D. Comparing the Ideation Quality of Humans with Generative Artificial Intelligence. IEEE Engineering Management. 2024. PP. 1-10. 10.1109/EMR.2024.3353338.

Kshetri N, Hutson, J Jayabaskar R. healthAIChain: Improving security and safety using Blockchain Technology applications in AI-based healthcare systems. 2023. 10.48550/arXiv.2311.00842.

Liu K, Zhang Y, Martin C, Ma X, Shen B. Translational Bioinformatics for Human Reproductive Biology Research: Examples, Opportunities and Challenges for a Future Reproductive Medicine. Int J Mol Sci. 2022 Dec 20;24(1):4. doi: 10.3390/ijms24010004. PMID: 36613446; PMCID: PMC9819745.

Muñoz E, Bronet F, Lledo B, Palacios-Verdú G, Martinez-Rocca L, Altmäe S, Pla J; representing the Special Interest Group in Reproductive Genetics of the Spanish Society of Fertility. To transfer or not to transfer: the dilemma of mosaic embryos - a narrative review. Reprod Biomed Online. 2024 Mar;48(3):103664. doi: 10.1016/j.rbmo.2023.103664. Epub 2023 Nov 2. PMID: 38408811.

Nassiri K, Akhloufi MA. Recent Advances in Large Language Models for Healthcare. BioMedInformatics. 2024 4(2), 1097-1143. https://doi.org/10.3390/biomedinformatics4020062

Ortiz JA, Morales R, Lledó B, Vicente JA, González J, García-Hernández EM, Cascales A, Ten J, Bernabeu A, Bernabeu R. Application of machine learning to predict aneuploidy and mosaicism in embryos from in vitro fertilization cycles. AJOG Glob Rep. 2022 Sep 19;2(4):100103. doi: 10.1016/j.xagr.2022.100103. PMID: 36275401; PMCID: PMC9574883.

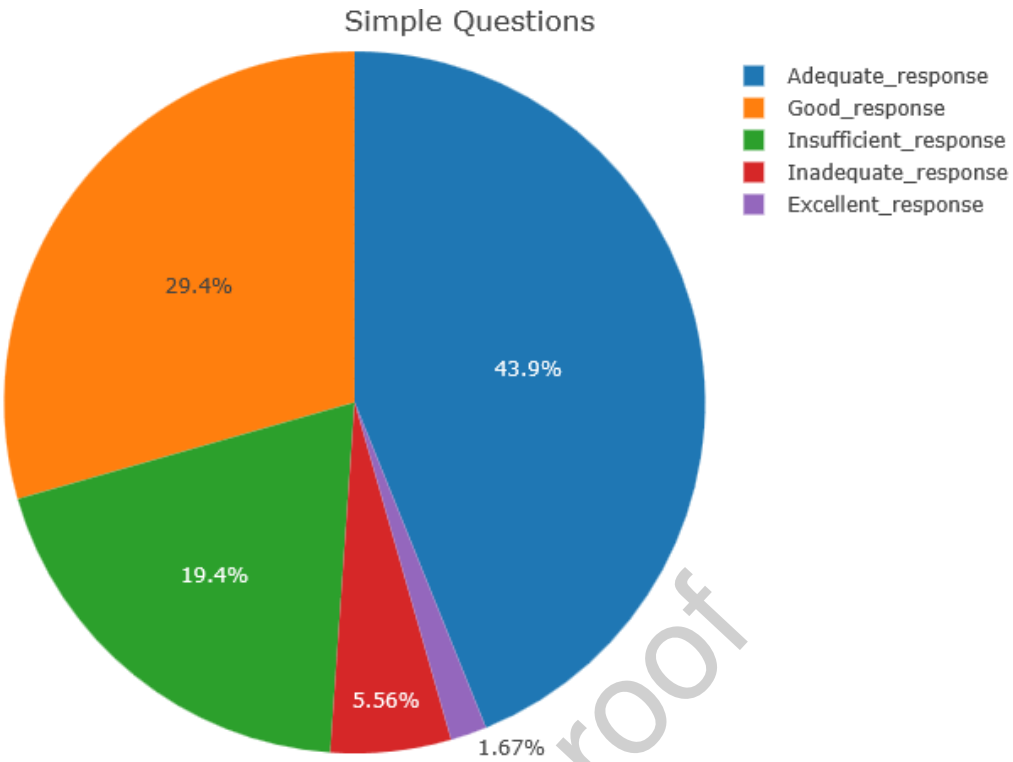
Reda, Menna Allah. (2024). Intelligent Assistant Agents: Comparative Analysis of Chatbots through Diverse Methodologies. 10.13140/RG.2.2.23344.57602.

Roumeliotis KI, Tselikas ND. ChatGPT and Open-AI Models: A Preliminary Review. Future Internet. 2023; 15(6):192. https://doi.org/10.3390/fi15060192

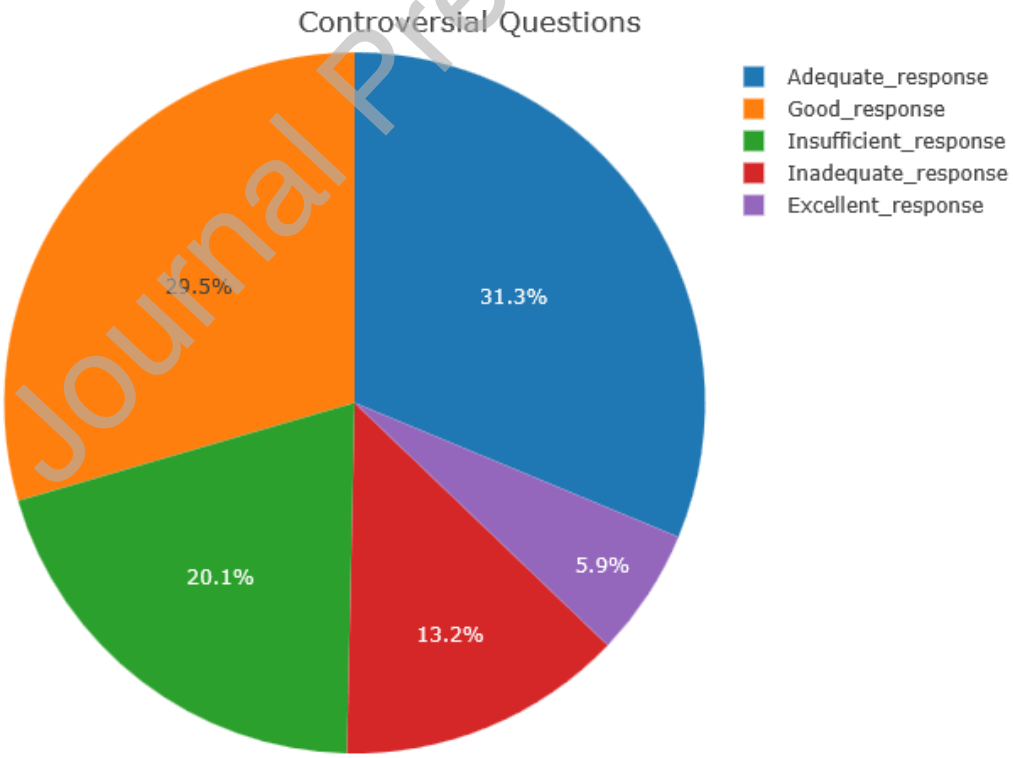
Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Roziere B., Goyal N. Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E., Lample G. LLaMA: Open and Efficient Foundation Language Models. 2023 10.48550/arXiv.2302.13971.

Schmidgall S, Harris C, Essien EI, Olshvang D, Rahman T, Kim JW, Ziaei R., Eshraghian J., Abadir P., Chellappa R. Evaluation and mitigation of cognitive biases in medical language models. 2024 npj Digit. Med. 7, 295. https://doi.org/10.1038/s41746-024-01283-6

Wang X, Ye H, Zhang S, Yang M, Wang X. Evaluation of the Performance of Three Large Language Models in Clinical Decision Support: A Comparative Study Based on Actual Cases. J Med Syst. 2025 Feb 14;49(1):23. doi: 10.1007/s10916-025-02152-9. PMID: 39948214.



1a



1b

Figure 1. The distribution of responses for (a) simple and (b) controversial questions

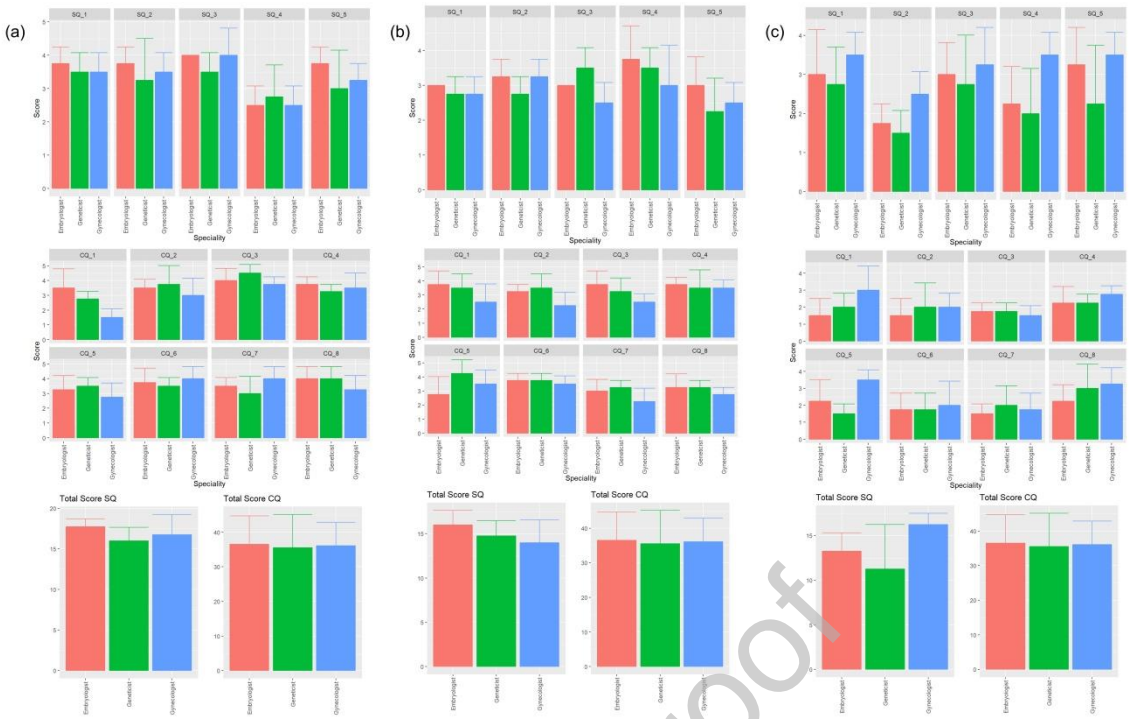


Figure 2. Comparative of the three chatbots by the panel expert (a) ChatGPT (b) Gemini (c) Llama

Table 1. Evaluation of three generative AI chatbots for Simple and controversial Questions regarding PGT and embryo mosaicism

	Total	ChatGPT [1]	Gemini [1]	Llama[1]	p-value
SQ_1	3.17 (0.70)	3.58 (0.51)	2.83 (0.39)	3.08 (0.90)	**0.022**[2]
SQ_2	2.83 (0.94)	3.50 (0.80)	3.08 (0.51)	1.92 (0.67)	**<0.001**[2]
SQ_3	3.28 (0.81)	3.83 (0.58)	3.00 (0.60)	3.00 (0.95)	**0.010**[2]
SQ_4	2.86 (0.96)	2.58 (0.67)	3.42 (0.90)	2.58 (1.08)	0.054[2]
SQ_5	2.97 (0.94)	3.33 (0.78)	2.58 (0.79)	3.00 (1.13)	0.107[2]
Total_Score_SQ	15.11 (2.87)	16.83 (1.80)	14.92 (2.02)	13.58 (3.60)	**0.0172**
CQ_1	2.67 (1.22)	2.58 (1.16)	3.25 (1.14)	2.17 (1.19)	0.088[2]
CQ_2	2.75 (1.18)	3.42 (1.00)	3.00 (0.95)	1.83 (1.03)	**0.004**[2]
CQ_3	2.97 (1.23)	4.08 (0.67)	3.17 (0.94)	1.67 (0.49)	**<0.001**[2]
CQ_4	3.17 (0.88)	3.50 (0.67)	3.58 (0.79)	2.42 (0.67)	**0.001**[2]
CQ_5	3.03 (1.13)	3.17 (0.83)	3.50 (1.17)	2.42 (1.16)	0.068[2]
CQ_6	3.08 (1.18)	3.75 (0.75)	3.67 (0.49)	1.83 (1.03)	**<0.001**[2]
CQ_7	2.69 (1.12)	3.50 (0.90)	2.83 (0.83)	1.75 (0.87)	**<0.001**[2]
CQ_8	3.22 (0.96)	3.75 (0.87)	3.08 (0.67)	2.83 (1.11)	0.050[2]
Total_Score_CQ	23.58 (6.51)	27.75 (4.49)	26.08 (3.99)	16.92 (4.96)	**<0.001**[3]

1 Mean (SD)

2 Kruskal-Wallis rank sum test

3 One-way ANOVA

SQ:Simple Questions
CQ:Controversial Questions

Table 2. Evaluation of three generative AI chatbots for Simple and controversial Questions regarding PGT and embryo mosaicism by specialist

	Geneticist				Embryologist				Gynecologist			
Characteristic	**ChatGPT [1]**	**Gemini [1]**	**Llama[1]**	**p-value**	**ChatGPT [1]**	**Gemini [1]**	**Llama[1]**	**p-value**	**ChatGPT [1]**	**Gemini [1]**	**Llama[1]**	**p-value**
SQ_1	3.50 (0.58)	2.75 (0.50)	2.75 (0.96)	0.253[2]	3.75 (0.50)	3.00 (0.00)	3.00 (1.15)	0.243[2]	3.50 (0.58)	2.75 (0.50)	3.50 (0.58)	0.153[2]
SQ_2	3.25 (1.26)	2.75 (0.50)	1.50 (0.58)	0.040[2]	3.75 (0.50)	3.25 (0.50)	1.75 (0.50)	0.013[2]	3.50 (0.58)	3.25 (0.50)	2.50 (0.58)	0.093[2]
SQ_3	3.50 (0.58)	3.50 (0.58)	2.75 (1.26)	0.526[2]	4.00 (0.00)	3.00 (0.00)	3.00 (0.82)	0.029[2]	4.00 (0.82)	2.50 (0.58)	3.25 (0.96)	0.088[2]
SQ_4	2.75 (0.96)	3.50 (0.58)	2.00 (1.15)	0.146[2]	2.50 (0.58)	3.75 (0.96)	2.25 (0.96)	0.088[2]	2.50 (0.58)	3.00 (1.15)	3.50 (0.58)	0.253[2]
SQ_5	3.00 (1.15)	2.25 (0.96)	2.25 (1.50)	0.577[2]	3.75 (0.50)	3.00 (0.82)	3.25 (0.96)	0.373[2]	3.25 (0.50)	2.50 (0.58)	3.50 (0.58)	0.093[2]
Total_Score_SQ	16.00 (1.63)	14.75 (1.71)	11.25 (4.99)	0.280[2]	17.75 (0.96)	16.00 (1.63)	13.25 (2.06)	**0.024[2]**	16.75 (2.50)	14.00 (2.58)	16.25 (1.26)	0.297[2]
CQ_1	2.75 (0.50)	3.50 (1.00)	2.00 (0.82)	0.099[2]	3.50 (1.29)	3.75 (0.96)	1.50 (1.00)	0.060[2]	1.50 (0.58)	2.50 (1.29)	3.00 (1.41)	0.252[2]
CQ_2	3.75 (1.26)	3.50 (1.00)	2.00 (1.41)	0.147[2]	3.50 (0.58)	3.25 (0.50)	1.50 (1.00)	0.040[2]	3.00 (1.15)	2.25 (0.96)	2.00 (0.82)	0.422[2]
CQ_3	4.50 (0.58)	3.25 (0.96)	1.75 (0.50)	0.015[2]	4.00 (0.82)	3.75 (0.96)	1.75 (0.50)	0.020[2]	3.75 (0.50)	2.50 (0.58)	1.50 (0.58)	0.013[2]
CQ_4	3.25 (0.50)	3.50 (1.29)	2.25 (0.50)	0.118[2]	3.75 (0.50)	3.75 (0.50)	2.25 (0.96)	0.035[2]	3.50 (1.00)	3.50 (0.58)	2.75 (0.50)	0.209[2]
CQ_5	3.50 (0.58)	4.25 (0.96)	1.50 (0.58)	0.016[2]	3.25 (0.96)	2.75 (1.26)	2.25 (1.26)	0.476[2]	2.75 (0.96)	3.50 (1.00)	3.50 (0.58)	0.380[2]
CQ_6	3.50 (0.58)	3.75 (0.50)	1.75 (0.96)	0.030[2]	3.75 (0.96)	3.75 (0.50)	1.75 (0.96)	0.037[2]	4.00 (0.82)	3.50 (0.58)	2.00 (1.41)	0.114[2]

CQ_7	3.00 (1.15)	3.25 (0.50)	2.00 (1.15)	0.264[2]	3.50 (0.58)	3.00 (0.82)	1.50 (0.58)	0.024[2]	4.00 (0.82)	2.25 (0.96)	1.75 (0.96)	0.035[2]
CQ_8	4.00 (0.82)	3.25 (0.50)	3.00 (1.41)	0.286[2]	4.00 (0.82)	3.25 (0.96)	2.25 (0.96)	0.080[2]	3.25 (0.96)	2.75 (0.50)	3.25 (0.96)	0.555[2]
Total_Score_CQ	28.25 (2.87)	28.25 (4.65)	16.25 (5.12)	**0.005[3]**	29.25 (4.50)	27.25 (2.06)	14.75 (3.77)	<**0.001**[3]	25.75 (6.08)	22.75 (3.10)	19.75 (5.68)	0.302[3]

[1] Mean (SD)

[2] Kruskal-Wallis rank sum test

[3] One-way ANOVA

She received her PhD in molecular biology at the University of Alicante, Spain. In 2004 she moved to Instituto Bernabeu. Nowadays, she is the Director of the Molecular Biology Department. She has received some prizes in different congress and has published tens of papers focus on genetic variants in infertility and preimplantation genetic diagnosis.

Generative AI shows moderate reliability in PGT-related queries, with ChatGPT and Gemini outperforming Llama. Performance varies by question complexity, highlighting the need for expert oversight and further refinement before clinical use in reproductive medicine.